Information Filtering Using SVD and ICA

Takeru Yokoi
Computer and Systems Science
Osaka Prefecture University
1-1 Gakuencho Sakai-City
Osaka 599-8531
takeru@sig.cs.osakafu-u.ac.jp

Hidekazu Yanagimoto
Computer and Systems Science
Osaka Prefecture University
1-1 Gakuencho Sakai-City
Osaka 599-8531
hidekazu@cs.osakafu-u.ac.jp

Sigeru Omatu
Computer and Systems Science
Osaka Prefecture University
1-1 Gakuencho Sakai-City
Osaka 599-8531
omatu@cs.osakafu-u.ac.jp

Abstract

We propose an information filtering system using Singular Value Decomposition(SVD) and Independent Component Analysis (ICA). The number of the independent components to estimate increases as the number of documents increases. Therefore, ICA requires much calculation amount according to it. When ICA is applied to documents, it is considerd that topics included in the documents are obtaied. However, it is difficult to clearly recognize the meaning of topics obtained by ICA. To solve these problem, before applying ICA, we transform the documents with SVD. Using SVD, we expects accuracy of the topic extraction becomes better and effects a user profile well. Then, in our proposed method, we map the document vectors into topic space. We create the user profile from transformed documents with Genetic Algorithm. Finally, we recommend documents by the user profile and evaluate accuracy by 11-point average precision. We carry out an experiment to confirm advantage of the poposed method.

Keywords: Independent Component Analysis, Singular Value Decomposition, Information Filtering System, User Profile

1 Introduction

As information technologies have been advanced, a plenty of information are served in the Internet. It has been difficult to find what we demand from large amount of information by existing retrieval systems since we cannot express our query exactly. A lot of researchers pay attention to developing information retrieval systems which automatically select the information depending on our interests.

The information retrieval systems with user's interests have been studied. For example, there are ranking methods that sort information depending on the user's interests and filtering systems which select the information depending on the user's interests. It is reported the method such as Latent Semantic Analysis(LSA)[1]. LSA is the method which chooses axes to focuse on the variance and cuts off noise. In LSA, Singular Value Decompositon(SVD) is often used to search the axes.

We search the axes from the viewpoint of independence instead of variance. Independent Component Analysis(ICA) is the method to find the axes depending on the independene. ICA is recently used for signal processing, image processing and so on. It has been already reported that we can obtain topics included in the documents on applying ICA to documents[2][3]. We use the topics for transformation of document vectors and improve recommendation accuracy of documents.

However, it is difficult to clearly understand meaning of some topics obtained by ICA. This is why input documents includes some noise. Hence, we use SVD to cut off the noise. It is reported to use SVD before applying ICA, for the accuracy of topic extraction to become better[4].

In this paper, we combine denoising by SVD and topic extraction by ICA to improve the user profile. Then we map the document vectors into the space which consists of the topics and construct the user profile with the transformed document vectors by GA. Finally, to confirm the proposed method, we carry out an experiment on test collection (NTCIR2[5]). In addition, we discuss about the topics obtaind by this algorithm.

2 Our proposed method

In this chapter, we explain a user profile, ICA and SVD.

2.1 User Profile

A document vector is a row vector whose elements are weights of words in a document. When the number of words is n and the weight for the ith word is w_i, the document vector x is denoted as

$$x = [\mathbf{w}_1 \quad \mathbf{w}_2 \quad \cdots \quad \mathbf{w}_n]^T \tag{1}$$

where $[\cdot]^T$ means transportation.

A user profile is also denoted by a row vector whose elements are weights of words like a document vector. In the user profile, interesting words need to have large weights and uninteresting ones need to have small weights. We construct the user profile using Genetic Algorithm(GA). In order to create a user profile which fulfills the conditions mentioned above, the fitness function of GA is defined as

$$f = \sum_i \alpha_i G^T D_i \tag{2}$$

where G means a gene, D_i means a document and α_i is a coefficient for each document.

2.2 Abstract of SVD

SVD is a method to look for an axis of large variance. It is reported that when SVD is applied to documents, we can analysis the word cooccurrence[1].

Now, we assume that m document vectors denoted as $x_1, x_2, \cdots, x_m$ and a document vector matrix X is denoted by equation(3).

$$X = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \cdots \quad \mathbf{x}_m]^T \tag{3}$$

Then SVD decomposes X into the equation(4), using two orthogonal matrices U, V and one diagonal matrix Σ.

$$X = U\Sigma V^T \tag{4}$$

The row vectors of U are called singular vectors and orthogonal bases of X. In this paper, we select k singular vectors with a large singular value to remove noise. In addition, k is determined by the rate of contribution. U_k represents these k singular vectors. The input matrix X is changed into X_k as follows using U_k.

$$X_k = U_k^T X \tag{5}$$

2.3 Extraction of Topics Using ICA

In signal processing, ICA extracts independent signals from some mixed signals. When ICA is applied to speech processing, observed variables are time series data recorded by microphones and independent variables are source signals. On the other hand, when ICA is applied to documents, the inputs of microphones correspond to document vectors.

Now m independent components are denoted as $s_1, s_2, \cdots, s_m$ and independent component matrix S are denoted by equation(6).

$$S = [\mathbf{s}_1 \quad \mathbf{s}_2 \quad \cdots \quad \mathbf{s}_n]^T \tag{6}$$

At that time, X_k, which is mensioned previous section, is assumed to follow a formula(7).

$$X_k = AS. \tag{7}$$

Here, A is an m k full rank mixing matrix. In addition, we assume $m \geq k$. If A is known, we can obtain the generalized inverse matrix $A^\dagger$ of A easily. However, $A^\dagger$ cannot be generally found because the mixing matrix A is unknown.

The purpose of ICA is to estimate a topic matrix S with only observed variables X_k under the condition where under the condition where mixing matrix A is unknown. In the other word, ICA finds a restored signal matrix Y which is statistically independent using the restored matrix W in the following equation.

$$Y = WX_k. \tag{8}$$

In addition, by the property of evaluation criteria, a magnitude and an order of the restored signals have not been determined uniquely.

Fast ICA[6] is one of ICA solution algorithms. This paper uses Fast ICA to find the independent components. The update criteria by hyperbolic tangent to find the independent components is equation(9).

$$\begin{aligned} w^+ &= E[Yg(w^TY)] - E[g'(w^TY)]w \\ g(u) &= tanh(u) \end{aligned} \tag{9}$$

In this paper, A which is inverse matrix of W is used as a feature axis showing a topic included in the documents. In fact, the topic is denoted by equation(10) since space conversion by U_k is performed before applying ICA.

$$Topic = U_k A \tag{10}$$

A Document vectors $\mathbf{x_i}$ is mapped to the space constructed by topics and represented with the topics. Here, we construct a user profile with $\hat{X}$ represented by topics in equation(11).

$$\hat{X} = Topic * X^T. \tag{11}$$

3 Experiment and result

3.1 Experiment environment and procedure

The data for an experiment are the 625 documents concerning with information retrieval from test collection NTCIR2. These documents have already been evaluated whether each document is relevant or not. In the documents, there are 34 relevant documents.

Each document is represented as a vector with vector space model[7]. As a methodology to represent a document with a vector, at first, we apply morphological analysis tool ChaSen[8] to documents and extract nouns. After that, we remove stop words and high frequency words thorough all documents. We set the threshold of frequency with 20 documents. With the above process, we get 5,948 words and the dimension of document vector is 5,948. Using tf-idf, these words are weighted.

We apply SVD to 625 document vectors and pick up axes until these contribution ratio is 0.8. Then we obtain 409 unique vectors. The input document vectors are changed with these vectors and we apply ICA to them. After that, we converted input documents with topics obtained by ICA.

We use GA for construction of a user profile in cross validation. We provide 625 documents into 5 subsets which include 125 documents. The number of relevant documents and non-relevance ones included in each subset is showed in Table 1. We put 3 subsets together as training data for the construction of the user profile and set the others with evaluate data. Hence, we carry out experiments on 10 patterns of training data.

In GA, each element of individual is expressed with 5 bit. To make the ratio of relevant documents and non-relevance ones set to 1:1, the coefficient α_i in the equation(2) is defined as

$$\alpha_i = \begin{cases} +1 & Di : Relevant \\ -N_I/N_J & Di : Non-relevance \end{cases}$$

where N_I means the number of relevant documents and N_J means the number of non-relevance documents. Other parameters of GA are shown in Table 2 and we use two point crossover. In addition we construct the user profile at 5 times and the average of 5 times results since the GA is a probabilistic approach.

Table 1: The Input Data.

@@@@	all	set1	set2	set3	set4	set5
relevant	34	7	13	5	3	6
non-relevance	591	118	112	120	122	119

Table 2: The Parameter of GA

@@@@	Generation	Cross Over	Mutation
Only GA	10000	1	0.005
ICA+GA	10000	1	0.05
SVD+ICA+GA	10000	1	0.05

Finally, we recommend evaluation documents depending on the user profile and evaluate accuracy of recommendation with 11-point average precision ratio. The recommended documents are determined depending on the similarity S_i between the user profile and the ith document vector D_i. Similarity S_i is defined as

$$S_i = U^T D_i \tag{12}$$

where U means user profile.

We summarize the experiment in the following steps.

Step1 Make document vectors with vector space model.

Step2 Apply SVD to document vectors and convert the space of input documents with unique vectors.

Step3 Apply ICA to converted documents.

Step4 Return the space of A with unique vectors and it is defined topics.

Step5 Transform the input documents with topics.

Step6 Construct the user profile with Genetic Algorithm.

Step7 Recommend documents and evaluate the user profile with 11-point average precision ratio.

Moreover, we construct the user profile with other 2 methods to confirm the advantage of the proposed method, which are construction of the user profile using only GA with original documents and ICA and GA without SVD. The parameters of GA used in these comparable experiments are shown in Table 2.

3.2 Results

In this section, we show the result of the evaluation experiment. Figure 1 shows recall precision curves and Table 3 shows 11-point average precision ratios. In addition, Table 4 shows the feature axes. Table 4 shows the some words sorted by absolute of word weights.

Figure 1: Imputation Precision with GA.

Table 3: 11-points Average Precision Ratio

	Original	ICA	ICA+SVD
0	0.160	0.322	0.331
0.1	0.133	0.218	0.294
0.2	0.126	0.139	0.219
0.3	0.121	0.115	0.171
0.4	0.118	0.099	0.145
0.5	0.116	0.096	0.133
0.6	0.112	0.089	0.124
0.7	0.108	0.082	0.113
0.8	0.106	0.075	0.103
0.9	0.101	0.070	0.087
1	0.098	0.067	0.076
Average	0.118	0.125	0.163

4 Discussion

From the Figure 1, the precision ratio of ICA version becomes better than original version. This is the reason why the concretization of topics contributes the improvement of precision ratio. The proposed method which uses SVD also left much better result than only ICA at low recall ratio especially. In addition, comparing the result from the viewpoints of 11-points average ratio in Table 3, the proposed method left good average precison. According to this, it can be said the proposed method has realized the totally improvement. This is because the accuracy of the topic used for conversion went up, which is shown in Table 4. Seeing Table 4, we can estimate the topics clealy. It is considered that Axis1 means "Cross Language Retrieval" and Axis2 means "Processor Array".

5 Conclusion

In this paper, we proposed the method of improving the accuracy of the user profile by detection of correct topic. Then, we confirmed the advantage of information filtering accuracy. Consequently, it checked that a user profile became good by improvement of topic detection.

Table 4: Example of Feature Axes

Axis1	Axis2
Language crossing	Array
Information access	Processor array
Use field	Processor
List stage	Mapping
Document selection	Physical array
Translation display	Routing
Real use	Size
Technical know-how	Failure processor
International exchange	Logic array
Gate	Mapping algorithm

References

[1] Deerwester.S, Dumais.T, Landauer.T, Furnas.W, Harshman.A, "Indexing by Latent Semantic Analysis", *Journal of the Society for Information Science* Vol.41, No.6, pp.391–497.

[2] Ata Kabán and Mark Girolami, "Topic Separation and Keyword Identification in Document Collections : A Projection Approach", Technical Report available in http://cis.paisley.au.ck/reseach/reports/index.html.

[3] T.Kolenda and L.K.Hansen, "Independent Components in Text", Advances in Independent Component Analysis, Springer-Verlag, 2000.

[4] Masafumi Hamamoto, Hiroyuki Kitagawa, Jia-Yu PAN and Christos Faloutsos, "Topic Detection from Text Data Using Independent Component Analysis", DEWS2004 3-B-04

[5] "NTCIR2", NII-NACSIS Test Collection for IR System, http://research.nii.ac.jp/ntcir/index-en.html.

[6] Aapo Hyvārinen, Erkki Oja, "Independent component analysis: A tutorial", Neural Network, Vol. 13, pp. 411–430, 2000.

[7] G.Salton, M.J.McGill, "Introduction to Modern Information Retrieval", McGraw-Hill Book Company, 1983.

[8] Y.Matsumoto, "Japanese Morphological Analysis System:CHASEN", Information Science Technical Report NAIST-IS-TR97007, Nara Institute of Science Technology,1997.